EVALUATING FORECASTERS' RULES OF THUMB :

A STUDY OF D(PROG) / DT

Thomas M. Hamill

University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

3 April 2003 (revised)

submitted as a NOTE to *Weather and Forecasting*

Corresponding author address: Dr. Thomas M. Hamill, NOAA-CIRES CDC, R/CDC 1, 325 Broadway, Boulder, CO 80305-3328. e-mail: tom.hamill@noaa.gov

ABSTRACT

Forecasters often develop rules of thumb for adjusting model guidance. Ideally, before use, these rules of thumb should be validated through a careful comparison of model forecasts and observations over a large sample. Practically, such evaluation studies are difficult to perform because forecast models are continually being changed, and a hypothesized rule of thumb may only be applicable to a particular forecast model configuration.

A particular rule of thumb was examined here, "$d$prog/dt". Given a set of lagged forecasts from the same model all verifying at the same time, this rule of thumb suggests that if the forecasts show a trend, this trend is more likely than not to continue and thus provide useful information for correcting the most recent forecast. Forecasters may also note the amount of continuity of forecasts to estimate the magnitude of the error in the most recent forecast.

Statistical evaluation of this rule of thumb was made possible here using a data set of forecasts from a "frozen" model. A 23-year record of forecasts was generated from a T62 version of the medium-range forecast model used at the National Centers for Environmental Prediction. Forecasts were initialized from reanalysis data, and January-February-March forecasts were examined for selected locations. "$d$prog/dt" was evaluated with 850 hPa temperature forecasts. A total of 2070 sample days were used in the evaluation.

Extrapolation of forecast trends was shown to have little forecast value. Also, there was only a small amount of information on forecast accuracy from the amount of discrepancy between short-term lagged forecasts. The lack of validity of this rule of thumb suggest that others should also be carefully scrutinized before use.

1. INTRODUCTION

Numerical weather prediction (NWP) models grow increasingly sophisticated with each passing year. Unfortunately, the quest for an NWP model free of systematic error remains elusive. Weather forecasters often develop rules of thumb to adjust the guidance produced by NWP models. Sometimes a rule of thumb may become obvious from a small sample. If, say, Eta model (Black 1994, Rogers et al. 1994, 1995, Mesinger 1996) forecasts are consistently too cold over snow every day for a month, a forecaster would certainly be wise to compensate for this bias until the model is improved. Nonetheless, human forecasters are fallible; their rules may appear to be appropriate from a relatively small sample of recent forecasts, but human judgment can often be a poor arbiter of statistical significance (see Gilovich 1993 for some interesting examples). Ideally, a forecaster should validate statistically their rules of thumb with a longer time series of forecasts. Practically, this takes time and effort, and a statistically robust sample may not be available, since operational weather prediction centers frequently update their weather forecast models,

Given these model changes, rules of thumb for adjusting model forecasts that can be applied regardless of the specific forecast model would be especially valuable. One potentially fruitful avenue for improving upon the latest numerical guidance is to consider multiple forecasts from the same model valid at the same time. Such *lagged-average forecasts* ("LAFs"; Hoffman and Kalnay 1983, Dalcher et al. 1988) have previously been shown to be useful for improving the skill of medium-range forecasts. For shorter-range forecasts, an evaluation of trends in lagged forecasts is often referred to informally as "$d\mathrm{prog}/dt$." Thus, one may see in forecast discussions that "temperature $d\mathrm{prog}/dt$ is negative," meaning that more recent numerical forecasts are colder than older ones. Some forecasters may also view $d\mathrm{prog}/dt$ as a handy, model- independent rule of thumb: if forecasts are trending colder, does that not suggest that the most likely actual state is yet somewhat colder than the most recent forecast? Forecasters may also note the amount of continuity of these lagged forecasts as a judge of the likely error in the most recent forecast.

3

Lagged forecasts that have been consistent are judged to be more accurate than ones that substantially differ from each other.

Because $d\mathrm{prog}/dt$ is often used as a rule of thumb regardless of the forecast model, it should be generally valid and testable with almost any model that can be run long enough to generate a statistically significant sample. Ideally, the most appropriate data to test would be the ones forecasters are now using. Hence, if forecasters are applying $d\mathrm{prog}/dt$ to 12, 24, and 36-h forecasts from the Eta model, this model should be tested. However, the Eta model is frequently modified at the National Centers for Environmental Prediction (NCEP), so a long history of forecasts from the current version of this model is not available. Consequently, we will test the validity with a forecast model where we do have a long time series of forecasts from the same model, a reduced-resolution version of NCEP's MRF model. If $d\mathrm{prog}/dt$ cannot be validated here, its applicability to more complex models should be considered suspect until demonstrated statistically.

The data to test $d\mathrm{prog}/dt$ were generated at the NOAA-CIRES Climate Diagnostics Center (CDC) in our "re-forecasting" project (http://www.cdc.noaa.gov/∼jsw/refcst). This project was undertaken in part to study whether significant improvements to forecast skill are possible if a very long time series of forecasts is available from a frozen model. Using this large training data set, systematic model errors can be detected, and current forecasts using the same frozen model can be adjusted for these errors. We have thus far generated 23 years of medium-range weather forecasts from a T-62 resolution version of the National Center for Environmental Prediction's (NCEP's) medium-range forecast (MRF) model (Kanamitsu 1989, Kanamitsu et al. 1991, Caplan et al. 1997, Wu et al. 1997). A single control forecast has been run forward for 2 weeks once every day from 0000 UTC initial conditions using the NCEP-NCAR reanalyses (Kalnay et al. 1996) from 1979 to 2001. By publishing time, we expect have completed a 15-member ensemble of forecasts over the 23 years. The reduced, T62 resolution was chosen so that the experiments could be conducted on the limited computer resources available at CDC.

The rest of the note consists of a brief examination of the skill of this forecast data set, an examination of how much improvement can be obtained through lagged regression approaches,

and an examination of the validity of the $d\text{prog}/dt$ rules of thumb. We hope the reader will see beyond the specifics of testing $d\text{prog}/dt$; the more important point is the importance of careful statistical evaluation of hypothesized rules of thumb.

2. RESULTS

Our data set will consist of 1, 2, and 3-day forecasts of 850 hPa temperature from January-March 1979-2001. Sea-level pressure forecasts were also examined but will not be shown here; the results were both qualitatively and qualitatively similar. NCEP-NCAR reanalyses were used as verification data. For simplicity, regression corrections and the usefulness of $d\text{prog}/dt$ was evaluated at a limited set of locations in the U.S.. These locations were the grid points nearest to Seattle, WA, Los Angeles, CA, Denver, CO, Minneapolis, MN, San Antonio, TX, Columbus, OH, Tampa, FL, Cape Hatteras, NC, and Portland, ME. To minimize the direct effect of forecast bias and the annual cycle upon the analysis, a 31-day running mean climatology of the analysis state and the mean forecast state was computed for each of these locations using the full 23-year data set. These running means were subtracted from the analyses and forecasts prior to the subsequent examination.

a. Validity of extrapolating forecast trends

First consider the overall error statistics of these forecasts. Table 1 provides the root mean-square (RMS) error characteristics of the forecasts at the nine locations as a function of lead time.

As a baseline for evaluating the value of forecast trends, a simple univariate regression was performed to predict the 850 hPa temperature provided from just the 24-h forecast temperature. For this regression, a cross-validation approach was used (Wilks 1995). The regression constants are separately calculated for each of the 23 years, using the remaining 22 years as training data. Denote T_{pred} as the predicted 850 hPa temperature (deviation from observed climatology) and T_{24} the 24-h forecast (deviation from forecast climatology). The regression equation was of the

form

$$T_{pred} = \beta_0 + \beta_1 \cdot T_{24} \tag{1}$$

The RMS error of this univariate regression is also displayed in column 5 of Table 1. The errors are consistently slightly lower than those from the 24-h forecast itself. On average, there was an $\sim$0.07 K reduction in RMS error.

If there is value in the trend in lagged forecasts, inclusion of these trends ought to significantly improve the accuracy of these forecasts. Accordingly, denote $(T_{48} - T_{24})$ the trend between 48 h and 24 h lagged forecasts valid at the same time, and similarly for $(T_{72} - T_{48})$. A cross-validated multivariate linear regression was performed of the form

$$T_{pred} = \beta_0 + \beta_1 \cdot T_{24} + \beta_2 \cdot (T_{48} - T_{24}) + \beta_3 \cdot (T_{72} - T_{48}). \tag{2}$$

The RMS errors of this multivariate regression are also displayed in the last column of Table 1. The inclusion of additional information on forecast trends made only a very small improvement to the skill of the forecasts; on average, only $\sim$0.02 K less than the errors from the univariate regression. If one examines the distribution of regression coefficients produced via the cross-validation (not shown), the distribution of β_2 and β_3 typically overlapped zero, indicating little confidence that the optimal values for these coefficients were significantly different from zero.

Examining a scatterplot of 48-24 h forecast trends and their relationship to the difference between the 24 h forecast and the analyzed state, the reason for the limited value of extrapolating trends is more apparent. Figure 1 provides this scatterplot; the difference in temperatures between 48 and 24-h forecasts valid at the same time is plotted along the x-axis, the difference between 24-h forecasts and the verification along the y-axis. There was little relationship between the forecast trend and the 24-h forecast error, as noted by the correlation coefficients near zero (plotted in the upper-left corner of each panel). The correlations were generally smaller yet if the trend was evaluated between 72 h and 24 h, and the correlations were no larger if one examined the subset of cases where there was a consistent trend in the 72-48 and 48-24 h forecast tendencies.

b. Estimating forecast skill from consistency

Is the consistency of forecasts useful for determining the accuracy of the most recent forecast? Figure 2 provides a scatterplot of the absolute difference between 48 and 24 h forecasts (x-axis) and the mean absolute error (MAE) of the 24 h forecasts (y-axis). Ideally, the larger the discrepancy between 48 and 24 h forecasts, the larger the typical MAE should be. Let F denote the absolute difference between 48 and 24 h forecasts. On Fig. 2, $\overline{MAE(F \leq 1)}$, $\overline{MAE(1 < F \leq 3)}$, and $\overline{MAE(3 < F)}$ is also plotted, where the overbar denotes the average over all forecasts. Note that there was only a small difference between the average MAEs of forecasts with large discrepancies and small discrepancies; the discrepancy in short-term lagged forecasts was only a slightly useful predictor of forecast skill.

3.   CONCLUSIONS

Weather forecasters develop rules of thumb to aid themselves in improving upon the numerical forecast guidance. Unfortunately, the human brain is often deceived into seeing patterns where there may be none (Gilovich 1993). Hence, a rule of thumb ought to be statistically validated before use, if this is possible. As an example of the potential problems with rules of thumb, we examined the usefulness of short-term lagged forecasts, i.e., $d$prog/dt. Using data from a reduced-resolution version of NCEP's MRF model and NCEP-NCAR reanalysis initial conditions, $d$prog/dt was shown to have little validity as a forecast rule of thumb. Short-term temperature trends with this model should not be extrapolated, and there is only a little value in the amount of discrepancy in lagged forecasts for predicting the magnitude of forecast error.

Is this apparent lack of improvement a consequence of using this particular model, or the better NCEP-NCAR initial conditions? While rules of thumb are often model-dependent, this particular rule seems to be applied across a variety of models and analysis systems. Following this same reasoning, $d$prog/dt should be carefully validated in other models rather than being used unquestioningly.

If not $d\mathrm{prog}/dt$, then what? There are demonstrably valuable techniques for estimating forecast uncertainty and improving the skill from a single deterministic forecast. One such technique is commonly referred to as *ensemble forecasting* (Toth and Kalnay 1993, 1997; Molteni et al. 1996; Houtekamer et al. 1996). There is a smaller body of literature on the usefulness of ensembles for shorter-range forecasts. See Brooks et al. (1992) for a motivation for short-range ensemble forecasting and Hamill et al. (2000) for a literature review. Other recent synoptic evaluations of ensembles include Mullen and Buizza (2001, 2002), Wandishin et al. (2001), and Grimit and Mass (2002). Though there are many challenging problems that need to be addressed to improve these forecasts, such data sets should be more useful for evaluating the uncertainty of shorter-range forecasts. Readers who may have used $d\mathrm{prog}/dt$ but are looking for a more theoretically justifiable alternative are encouraged to consider the information from these ensemble studies and to examine the new short-range ensemble forecast guidance now being generated at NCEP.

4. ACKNOWLEDGMENTS

REFERENCES

Black, T. L., 1994: The new NMC mesoscale eta model. Description and forecast examples. *Wea. Forecasting*, **9**, 265-284.

Brooks, H. E., C. A. Doswell, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 120-132.

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model analysis-forecast system. *Wea. Forecasting*, **12**, 581-594.

Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium-range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402-416.

Gilovich, T., 1993: How what we know isn't so: the fallibility of human reason in everyday life. Free Press, 216 pp.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific northwest. *Wea. Forecasting*, **17**, 192-205.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range. Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653-2664.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte-Carlo forecasting. *Tellus*, **35a**, 100-118.

Houtekamer, P. L. , L. Lefaivre, and J. Derome, 1996: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability, Vol II*, Reading, United Kingdom, 121-146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom].

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting*, **4**, 334-342.

——, and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6**, 425-435.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.*, **77**, 2637-2650.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638-663.

——, and ——, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173-191.

Rogers, E., D. G. Deaven, G. S. Dimego, 1995: The regional analysis system for the operational "early" Eta model: original 80-km configuration and recent changes. *Wea. Forecasting*, **10**, 810-810.

——, T. L. Black, D. G. Deaven, G. J. DiMego, Q. Zhao, M. Baldwin, N. W. Junker, Y. Lin, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting*, **11**, 391-416.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729-747.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences: an introduction.* Academic Press. 467 pp.

Wu, W. M. Iridell, S. Saha, and P. Caplan, 1997: Changes to the 1997 NCEP operational MRF model analysi/forecast system. *NCEP Technical Procedures Bulletin 443*, available from http://www.nws.noaa.gov/om/tpb/indexb.htm, or the National Weather Service, Office of Meteorology, Programs and Plans Division, Silver Spring, MD, 20910.

FIGURE CAPTIONS

Figure 1. Temperature trends ($d$prog/dt) and forecast errors for various locations. Plotted on x-axis in each panel are the 23 years × 90 days of 850 hPa temperature differences between 48-h and 24-h forecasts valid at the same time. On the y-axis are differences between the 24-h forecast and the verification. Correlation coefficient plotted in upper-left corner.

Figure 2. Mean absolute error (MAE) of 24-h 850 hPa temperature forecasts (y-axis) and their relationship to the absolute difference between 48-h and 24-h temperature forecasts (x-axis, forecasts valid at the same time). Horizontal lines denote the average MAE for cases when the absolute difference was less than 1 C, between 1 and 3 C, and greater than 3 C.
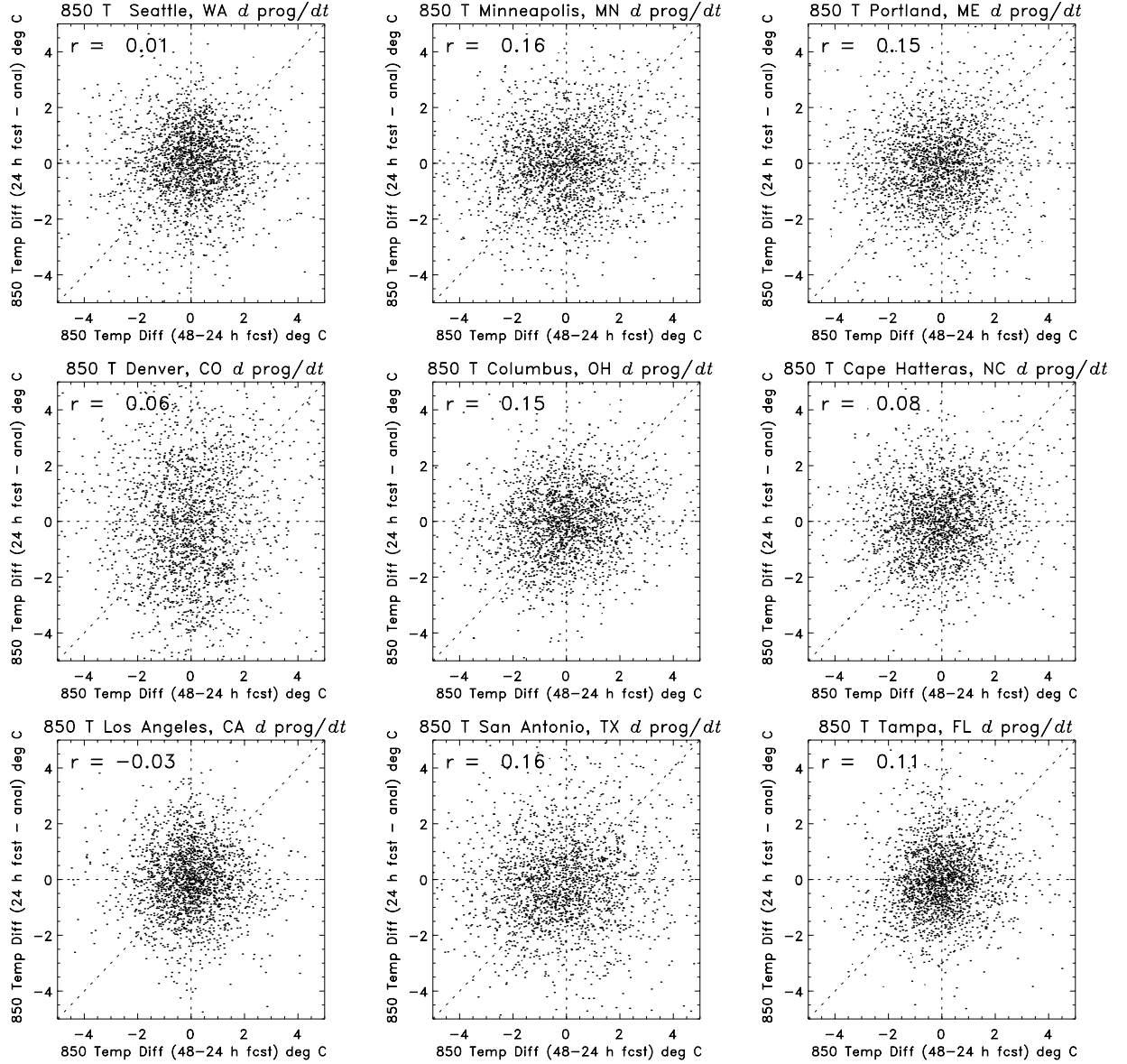
Figure 1. Temperature trends (*d*prog/*dt*) and forecast errors for various locations. Plotted on x-axis in each panel are the 23 years × 90 days of 850 hPa temperature differences between 48-h and 24-h forecasts valid at the same time. On the y-axis are differences between the 24-h forecast and the verification. Correlation coefficient plotted in upper-left corner.
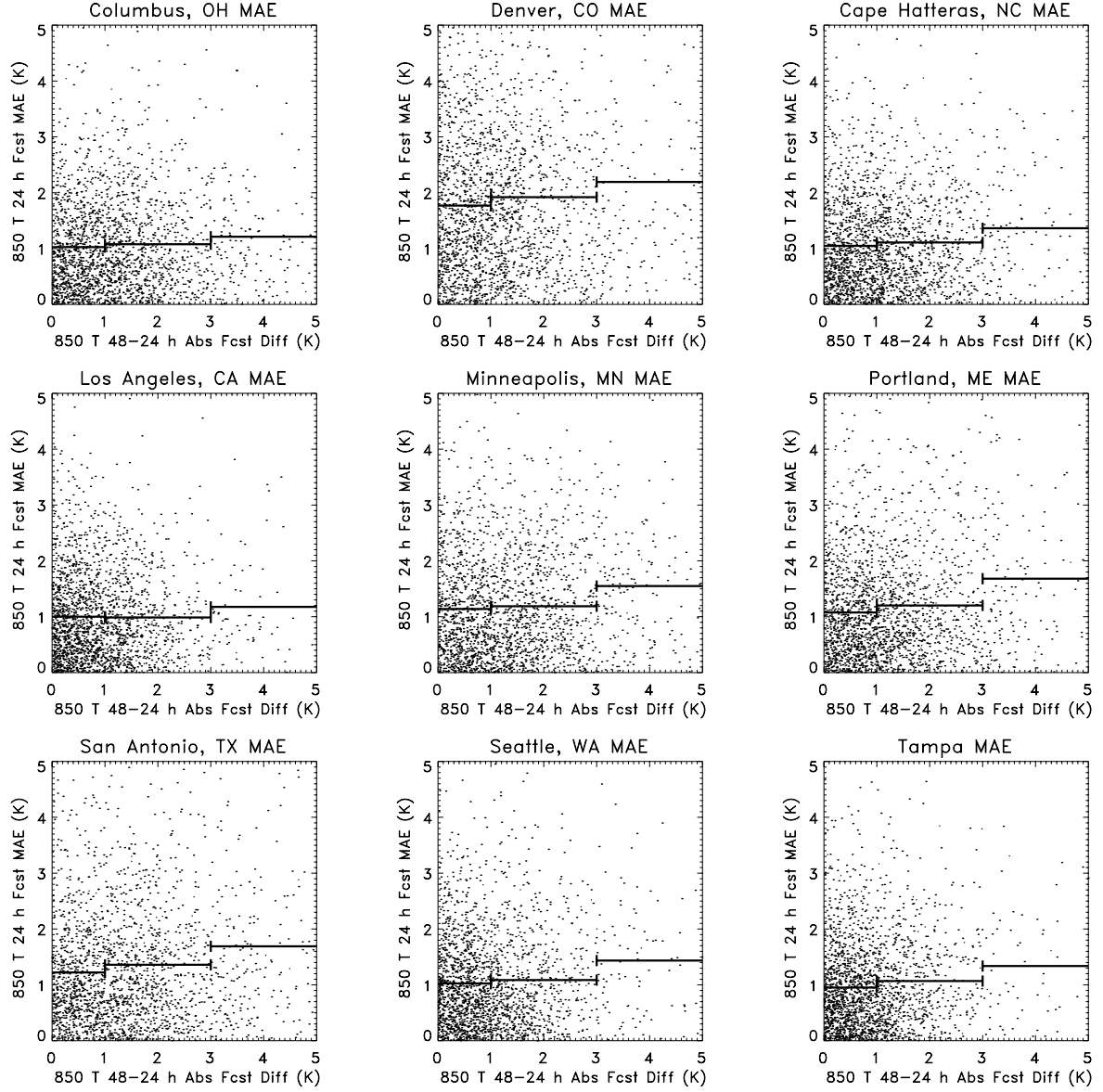
Figure 2. Mean absolute error (MAE) of 24-h 850 hPa temperature forecasts (y-axis) and their relationship to the absolute difference between 48-h and 24-h temperature forecasts (x-axis, forecasts valid at the same time). Horizontal lines denote the average MAE for cases when the absolute difference was less than 1 C, between 1 and 3 C, and greater than 3 C.

Table 1. 850 hPa RMS temperature errors (degrees C) for selected locations. First column de-
noted the location; columns 2, 3, and 4 denote the RMS errors of 24, 48, and 72 h forecasts,
respectively. Column 5 indicated the RMS error of a linear regression forecast with one pre-
dictor (24-h 850 hPa temperature), and column 6 the error of a multivariate linear regression
forecast with 3 predictors (24-h 850 hPa temperature, 48-24 h difference, and 72-48 h differ-
ence).

	24-h RMSE	*48-h RMSE*	*72-h RMSE*	*Regr1 RMSE*	*Regr3 RMSE*
Seattle, WA	1.38	2.02	2.59	1.33	1.32
Denver, CO	2.37	3.05	3.83	2.07	2.06
Los Angeles, CA	1.28	1.82	2.31	1.23	1.21
Minneapolis, MN	1.57	2.66	3.75	1.52	1.50
Columbus, OH	1.38	2.38	3.41	1.36	1.34
San Antonio, TX	1.80	2.90	3.82	1.72	1.67
Portland, ME	1.58	2.52	3.67	1.55	1.51
Cape Hatteras, NC	1.41	2.26	3.50	1.38	1.37
Tampa, FL	1.30	1.93	2.66	1.24	1.22